BITAG

Broadband Internet Technical Advisory Group

Latency Explained

Broadband Internet Technical Advisory Group

February 2022

Broadband Internet Technical Advisory Group

http://www.bitag.com

Twitter: @BITAGORG

Mailing Address:

1550 Wynkoop, Suite 168

Denver, CO 80202

Douglas C. Sicker

Executive Director

Chair of the Technical Working Group

BITAG

dsicker@bitag.org

720-724-8083

Greg White

Distinguished Technologist

CableLabs

g.white@cablelabs.com

Agenda

- Overview of BITAG
- Summary of the report
  - Observations
  - Recommendations
- Time for Q&A
  - Submitted and live questions

BITAG: *An Overview*

- **History** – Established in 2010 prior to the first Open Internet Order.
- **Mission** – BITAG brings together engineers and technical experts to develop consensus on broadband network management practices.
- **Technically Focused** – Technical Working Group (TWG) participants must meet technical requirements through education and/or experience.
- **Expeditious** – The TWG operates under a 120-day "shot clock" within which it must analyze the technical topic and generate a report.
- **Five Member Categories** – BITAG has five participating member categories designed to include all of the Internet ecosystem.
- **Balanced Processes and Consensus-Based Decision-Making** – The TWG strives to operate on a consensus basis, with backstop voting procedures.
- **Support for Independent Engineering Resources** – BITAG members value broad participation and provide support for independent engineers and academics to participate and represent the public interest.

BITAG Participants

- Shamim Akhtar, Apple
- Richard Bennett,HighTechForum
- Stuart Cheshire, Apple
- Sam Crawford, SamKnows
- Koen de Schepper, Nokia
- Amie Elcan, Lumen Technologies
- Nick Feamster, University of Chicago
- Cullen Jennings, Cisco
- Alan Jones, NetForecast
- Kate Landow, Dish
- Matt Larsen, Vistabeam
- Jason Livingood, Comcast
- Matt Mathis, Google
- David Reed, University of Colorado
- Peter Saint-Andre, Mozilla
- Alex Salter, SamKnows
- Kevin Schneider, Adtran
- Peter Sevcik, NetForecast
- Doug Sicker, BITAG
- Barbara Stark, AT&T
- Dave Täht, Teklibre
- Matt Tooley, NCTA
- Greg White, CableLabs
- David Winner, Charter Communications

Latency Explained

Motivation:

It is time to update our understanding of the primary factors directly affecting end-user internet performance. For over thirty years, industry and policymakers have collectively missed a key factor that drives end users' internet quality of experience (QoE).

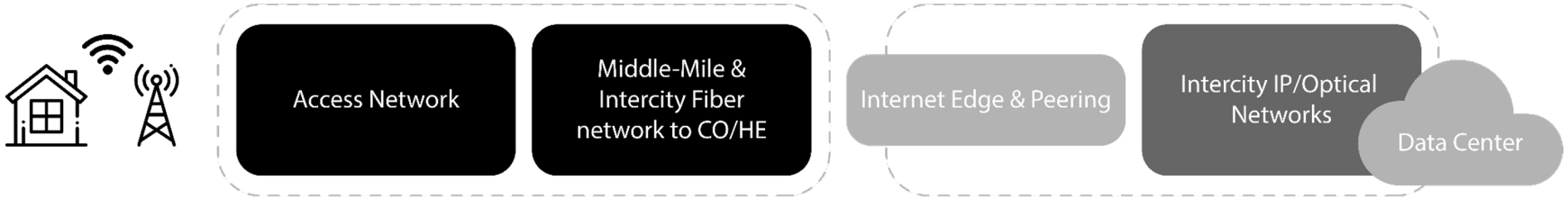
— Report published in January 2022

Topics covered by the report

- Definition of latency and its relationship to throughput/speed
- Impact that latency has on user experience for applications
- Sources/contributors to latency
- Metrics & methods for characterizing latency
- Current & future technologies to reduce latency
- Observations, findings and recommendations

What is network latency?

- **Defn**: The time it takes for a minimal data packet to travel from one network endpoint to another network endpoint.
  - The component of delay that doesn't vary with message size.
- A characteristic of the path between each pair of endpoints on the network.
  - Commonly varies over time.

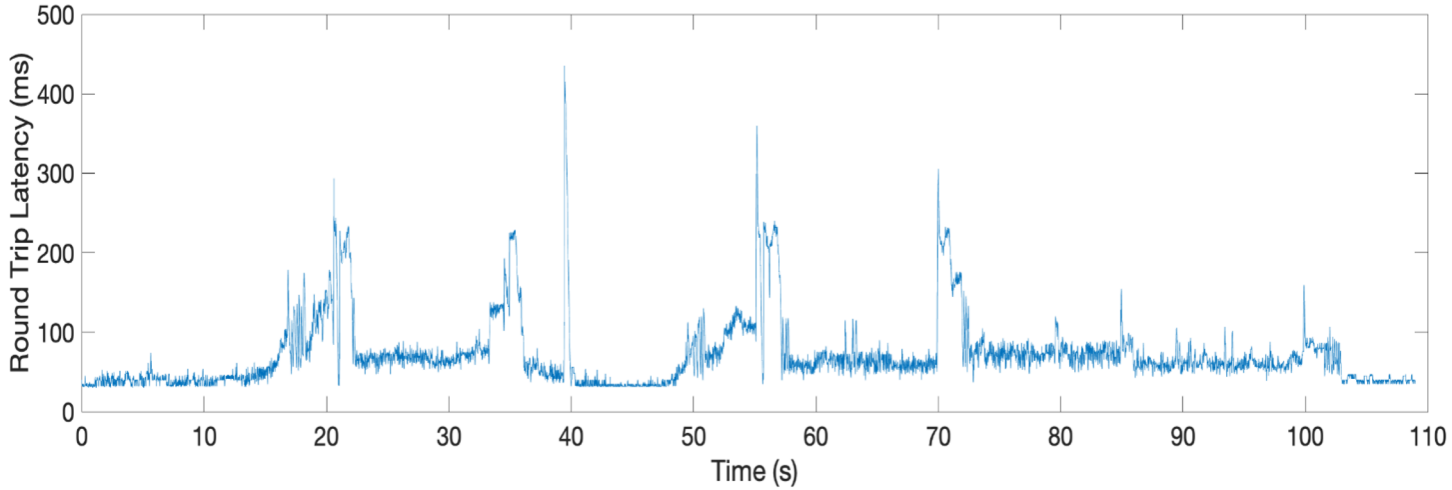


Observations & Findings

- The industry has historically been focused on increasing bandwidth, which has been straightforward, easy to understand and indisputably made dramatic improvements to end-user QoE.
- Latency is a critical factor in providing a high performance Internet connection. High latency negatively affects the quality of experience while using many applications.
- The way in which network latency has historically been characterized is very limited. The measurement methodology and the metrics typically used to describe latency have had very little to do with end-user QoE.
- “Working latency” is a better measurement of the end-user application QoE than idle latency.

Working Latency Over Time for a Network Connection

Historically, latency measurements have focused on average, idle latency.



Note: "Idle Latency" = 35 ms–this is not a useful metric.

Actual latency is over 100ms more than 10% of the time!

Spikes above 200ms about 4% of the time!

Many real-time applications deal with latency variation by buffering (delaying) early arriving packets to match the delay of the late arriving ones.

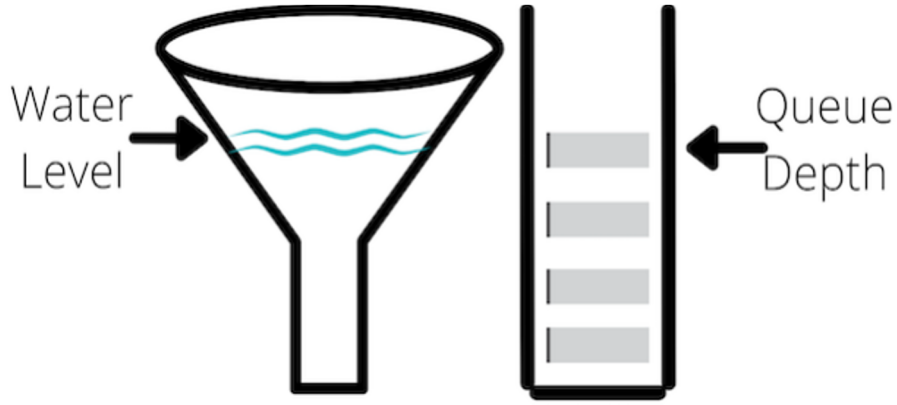
The only metric that matters is how high the peaks are!

Observations & Findings

- Working latency is valuable as it measures the real-world end-to-end user experience.
- Consensus on how to accurately and consistently measure working latency is still emerging.
  - But, it should focus on the peaks, not the average.

Observations & Findings

- One of the most impactful (and solvable) sources of latency and latency variation affecting Internet users is buffering delay.
  - Buffering delay arises from the interplay between application behavior in endpoints and queue management in the network.



Observations & Findings

- Queue management techniques such as Active Queue Management (AQM) are available that will reduce buffering delay in network equipment by triggering applications to reduce the amount of queuing delay that they cause.
- Very-low-latency networking technologies are emerging that aim to eliminate buffering delays altogether and seem likely to enable the creation of new classes of applications.

Recommendations

The industry should start to measure and report on working latency — in networks and in networking equipment — as this is often as critical to end-user experience as bandwidth capacity ("speed" or throughput).

- Continue to develop a testing method that accurately measures working latency;
- Highlight 98th or 99th percentile (or maximum) packet latency as the most salient metric;
- Don't report mean or median packet latency;
- Report on the variability by also including the minimum latency value.

Broadband internet access service providers and developers of network equipment should:

- Work to deploy mechanisms to reduce working latency.
- Investigate future methods for delivering very-low-latency services.

Recommendations

Application developers and operating system developers should:

- Investigate future methods for delivering very-low-latency services;
- Consider presenting working latency metrics to end-users in an easy-to-understand manner;
- Adapt application or operating system behavior in response to reductions in working latency;
- Use and promote the use of existing operating system features to eliminate excessive on-device buffering.

Policymakers should:

- Learn more about the evolving topic of working latency;
- Avoid creating barriers to the development and deployment of technologies that improve working latency**.**

BITAG

Broadband Internet Technical Advisory Group

Q&A Session

February 2022

BITAG

Broadband Internet Technical Advisory Group

Thank you!

February 2022